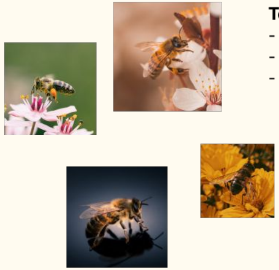
ViTopic – Topic Modeling on Images by Clustering of Vision Transformer Embeddings

By: Fausto J. German Jimenez

Problem Statement & Motivation

Image data plays a critical role in many fields such as social media analysis, medical imaging, and marketing. However, collecting, sorting, clustering, and labeling large amounts of image data can be a challenging and time-consuming task that often requires significant human resources. The proposed work aims to develop an automated system that can cluster and describe large datasets of unlabeled images, allowing analysts to quickly and accurately process vast amounts of visual data. This project's primary objective is to create a tool that can help people automatically cluster images into groups and generate text-based topics for each cluster, which will help them extract meaningful insights from the data. By automating this process, the system will significantly improve image data processing efficiency, reduce labor costs, and facilitate the analysis of visual data.

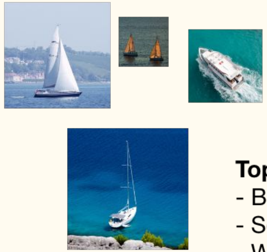
Main Idea – Visually Explained



Top 3 Descriptors:
- Bees
- Flowers
- Standing

Top 3 Descriptors:
- Birds
- Standing
- Surface

Main Idea
Each image is assigned a cluster based on its visual context (as determined by their vision transformer embeddings). Then, each image is described using an image captioning model, and text-based descriptors are generated based on the frequency of words in the generated captions.



Top 3 Descriptors:
- Boats
- Sea
- Waves

Top 3 Descriptors:
- Airplanes
- Sky
- Clouds

Reference Paper

Title: Contextual Visual Similarity

Authors: Xiaofang Wang, Kris M. Kitani and Martial Hebert

Year: 2016

Publisher: ArXiv

Reference Paper Continued

While I am not duplicating the work in the reference paper, I will use it as a background for the main idea behind my project, which is contextual visual similarity.

The main difference between the work in the reference paper and the methods I plan to use is that they require three images (A query image, a positive example, and a negative example) to define a contextualized similarity search criteria, while my project will compute similarity based on the Inverse Euclidean Distance between Vision Transformer Embeddings.

Data

I will use the ImageNette dataset, which is a small subset of the ImageNet project.

We can easily access this subset using the Datasets library from the HuggingFace repository: https://huggingface.co/datasets/frgfm/imagenette

This dataset will make it possible to evaluate the performance of the system since it contains labeled images within a small number of classes.

https://paperswithcode.com/dataset/imagenet

Survey of Related Work

Since there are very few papers that do exactly what I describe in this project, I have chosen three other papers that implement the main stages of the system I plan to build.

1. **Swin Transformer: Hierarchical vision transformer using shifted windows.**
    a. ByZe Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Published to ArXiv on August 17, 2021.
    b. **Why this paper?** It describes Vision Transformers, which I plan to use to compute image embeddings for contextual similarity.
2. **Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation**
    a. By Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Published to ArXiv on February 15, 2022.
    b. **Why this paper?** It describes the model I plan to use for automated image-captioning.
3. **BERTopic: Neural topic modeling with a class-based TF-IDF procedure.**
    a. By Maarten Grootendorst. Published to ArXiv on March 11, 2022.
    b. **Why this paper?** It describes the pipeline I plan to use to create contextual topics.

Summary of Methods

The reference paper uses triples of images for unsupervised attribute discovery, where they learn feature weights for each triplet such that the distance between two triplets is small if they are similar. In contrast, the method I plan to uses simply computes the pair-wise, inverse euclidean distance between embeddings generated from vision transformers.

Once embeddings are generated, I can use techniques similar to the ones described in “BERTopic: Neural topic modeling with a class-based TF-IDF procedure” to generate contextual topics from the image embeddings.

Then, using the “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation” model, I can generate captions for each of the images which will serve as topic descriptors.

Finally, using a class-based Term-Frequency Inverse-Document Frequency representation, I can generate keywords that are present in each of the image clusters, which will provide a text-based summary of the contextual topics.

Plan of Work

As daunting as the task sounds, breaking it into each of its constituent parts is easy. Within the next few weeks leading to the final delivery, I will work on each of the parts respectively. While it is impossible to determine the exact amount of time I will spend on each component of the system, I can provide a rough overview of how I plan to execute the project:

1. Week 1
   a. Read the referenced articles
   b. Prepare the dataset and pre-trained models
2. Week 2
   a. Combine the described methods into a coherent notebook demonstrating the project
   b. Test the system using the ImageNette dataset
3. Week 3
   a. Work on the presentation slides
   b. Work on the final report

References

- Grootendorst, M. (2022, March 11). BERTopic: *Neural topic modeling with a class-based TF-IDF procedure*. arXiv.org. Retrieved March 17, 2023, from https://arxiv.org/abs/2203.05794
- Li, J., Li, D., Xiong, C., & Hoi, S. (2022, February 15). *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*. arXiv.org. Retrieved March 17, 2023, from https://arxiv.org/abs/2201.12086Links to an external site.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021, August 17). *Swin Transformer: Hierarchical vision transformer using shifted windows*. arXiv.org. Retrieved March 17, 2023, from https://arxiv.org/abs/2103.14030Links to an external site.
- Wang, X., Kitani, K. M., & Hebert, M. (2016, December 8). *Contextual visual similarity. arXiv.org*. Retrieved April 1, 2023, from https://arxiv.org/abs/1612.02534Links to an external site.